Research and Application of Large Model-Based Intelligent Customer Service System

Yuan Xi [1]
[1]Beijing Ziwen Technology Co., Ltd.
xiyuan@vip.163.com

***Abstract*—With the rapid development of artificial intelligence technology, intelligent customer service systems have been widely used. This paper addresses the limitations of traditional intelligent customer service systems, such as limited language understanding ability, narrow knowledge coverage, and insufficient personalized service. It proposes an intelligent customer service system design scheme based on the RAG model. The scheme leverages the powerful language understanding and generation capabilities of large models, combined with dialogue management and knowledge base retrieval enhancement techniques, to build an efficient and intelligent customer service system. This paper introduces the overall architecture of the system, the design and implementation of each module, and comprehensively evaluates the system through experiments. The experimental results show that the system can provide accurate and fluent customer service, significantly improving customer satisfaction. The research in this paper provides new ideas and references for the development of intelligent customer service systems.**

***Index Terms*—Intelligent customer service, RAG large model, Dialogue management, Knowledge base, Modular design**

I. Research Background

As enterprises undergo digital transformation, customer service has become a key factor in enhancing competitiveness. Traditional manual customer service faces problems such as high labor costs, low efficiency, and unstable service quality. Intelligent customer service systems have emerged, empowered by artificial intelligence technology, to provide customers with high-quality service 24/7 [1]. In recent years, large model techniques, represented by the Transformer, have made significant breakthroughs and demonstrated excellent performance in the field of natural language processing. Large models have powerful language understanding and generation capabilities and can handle complex semantic relationships and contextual information [2]. Applying large models to intelligent customer service systems is expected to significantly improve the intelligence level and service quality of the systems.

II. Research Significance

Intelligent customer service systems are an important application of artificial intelligence technology in customer service and have broad development prospects. Exploring the design of intelligent customer service systems based on large models is of great significance for improving customer service automation, enhancing user experience, and reducing enterprise operating costs [3].

This research contributes to the technological progress and industrial development of intelligent customer service systems and provides enterprises with new solutions for better and more efficient customer service. At the same time, the research results also provide reference for intelligent applications in other fields.

A. Research Content and Methods

This paper designs and implements an intelligent customer service system with large model techniques at its core. The research content mainly includes:

1) Overall system architecture design: Design the overall architecture of the system and clarify the functions and interactions of each module.
2) Large model module design: Use the Transformer structure to build powerful language understanding and generation models.
3) Dialogue management module design: Design a dialogue state management mechanism to achieve smooth multi-turn dialogue interaction.
4) Knowledge base module design: Construct a domain knowledge base to support fast and accurate information retrieval.
5) System implementation and evaluation: Set up an experimental environment and conduct a comprehensive performance and user experience evaluation of the system.

The following research methods were used in this paper:

Literature research method: Comprehensively investigate the relevant literature on intelligent customer service and large models to understand the current state of technology development.

Modular design method: Adopt a modular design approach to improve the scalability and maintainability of the system.

Experimental evaluation method: Compare the performance of different models and systems through experiments to objectively evaluate the effectiveness of the system.

III. Related Work

A. Development History of Intelligent Customer Service Systems

Intelligent customer service systems have evolved from rule-based to intelligence-based. Early systems relied mainly on manually configured rules and templates, making it difficult to handle complex customer inquiries. In recent years, the

ISSN:3006-2985(Print)
ISSN:3006-9297(Online)

rise of deep learning technology has promoted the rapid development of intelligent customer service [4]. Researchers have explored building end-to-end dialogue systems using deep neural networks, enabling systems to automatically learn dialogue strategies and generate more natural and pertinent responses [5].

B. Research Progress in Large Model Technology

Large models are an important breakthrough in the field of natural language processing in recent years. Large models represented by the Transformer can acquire rich language knowledge and common sense through self-supervised learning on large-scale corpora [6]. GPT, BERT, XLNet, and other large models have achieved state-of-the-art performance on multiple tasks [7]–[9]. Researchers have also explored the application of large models in dialogue generation, question answering, summarization, and other tasks, achieving significant results [10], [11].

C. Application Status of Large Models in Intelligent Customer Service

Large models show enormous application potential in the field of intelligent customer service. Researchers use pre-trained models to build intelligent customer service systems, significantly improving the language understanding and generation capabilities of the systems [12]. Well-known intelligent assistants such as Microsoft Xiaoice and Google Meena have adopted large model technology [13], [14]. However, the application of large models in real customer service scenarios is gradually expanding. How to optimize the model training and inference process based on the characteristics of personalized customer service domains, solve the potential hallucination problem of large models, and achieve efficient and controllable customer service dialogue remains a pressing challenge [15].

IV. Design of Intelligent Customer Service System Based on Large Models

This section will introduce in detail the design of an intelligent customer service system based on RAG. The system adopts a modular design approach and mainly includes large model, dialogue management, and knowledge base modules. Through reasonable architecture design and technology selection, an efficient and intelligent customer service system is constructed.

A. Overall System Architecture Design

Figure 1 shows the overall architecture of the system.
Table I describes the functions of the system's core modules.

B. Large Model Module Design

The large model module is the core component of the system, responsible for semantic understanding and dialogue generation tasks. This system uses the Transformer encoder-decoder structure to build a large model. It acquires general language knowledge through self-supervised pre-training, and then fine-tunes on customer service domain data to achieve dialogue generation in specific domains [3], [16].

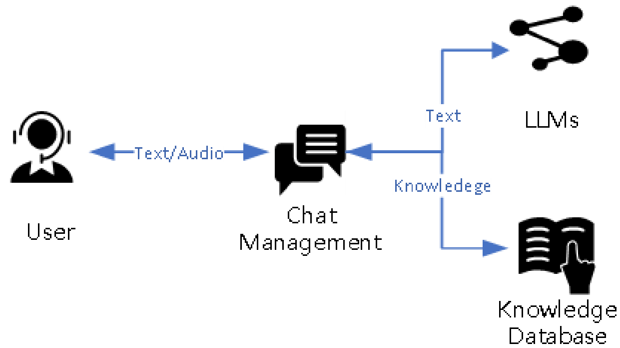

Fig. 1. Intelligent Customer Service System Overall Architecture

TABLE I
System Module Function Description

Module	Function Description
Dialogue Management Module	Responsible for receiving user input, managing dialogue state, and coordinating other modules to complete the dialogue process
Large Model Module	Responsible for core natural language processing tasks such as semantic understanding and dialogue generation, supporting dialogue generation
Knowledge Base Module	Stores domain knowledge and provides information retrieval functionality

Figure 2 shows a schematic diagram of the large model structure. The Transformer encoder is responsible for feature extraction and semantic encoding of user input, while the decoder autoregressively generates response text based on the encoding results and historical dialogue information.

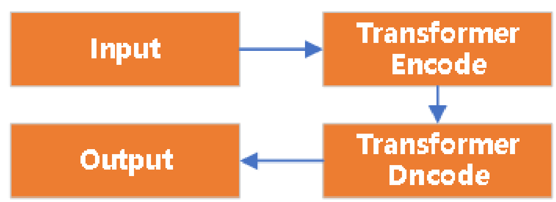

Fig. 2. Large Model Structure

Model training is divided into two stages: pre-training and fine-tuning. In the pre-training stage, self-supervised learning is performed on a large general corpus to acquire language knowledge through tasks such as Masked Language Modeling [6]. In the fine-tuning stage, supervised learning is performed on customer service domain data to optimize the model's performance in customer service scenarios [12]. Figure 3 shows the model training process.

C. Dialogue Management Module Design

The dialogue management module is responsible for controlling the dialogue process and managing multi-turn dialogue states. The system adopts a state machine-based dialogue management method, modeling the dialogue process as a series of transitions between states [4]. Each state represents a stage of the dialogue, and transitions between states are triggered

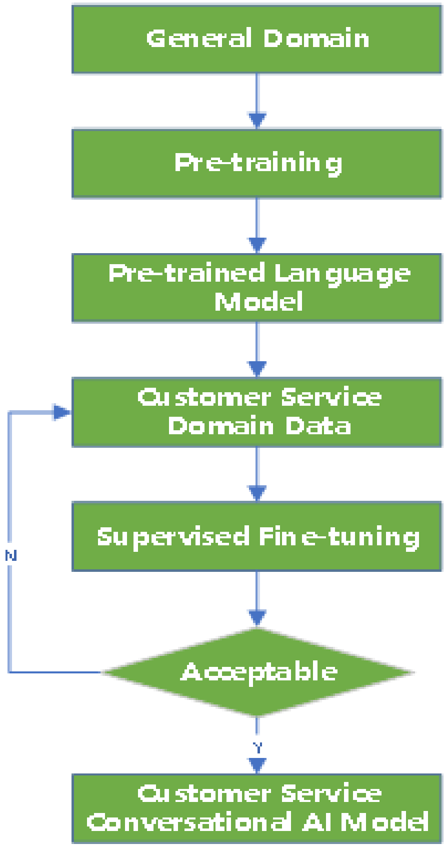

Fig. 3. Large Model Training Process

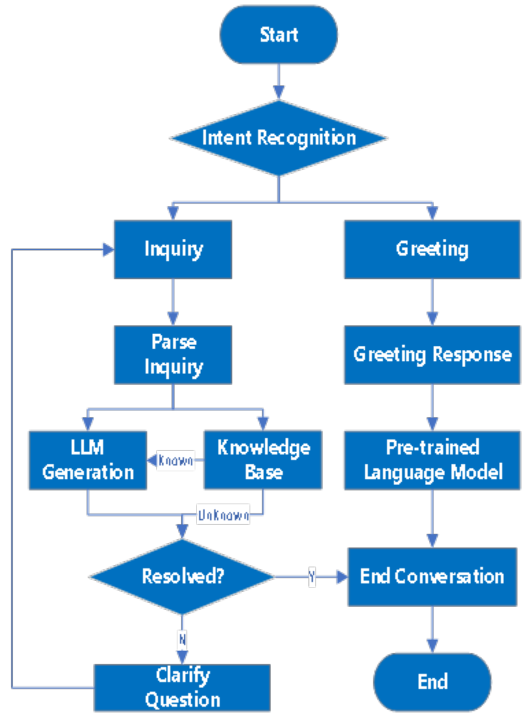

Fig. 4. Dialogue State Management Process

by recognizing user intent and slot values. Figure 4 shows the dialogue state management process.

D. Knowledge Base Module Design

The knowledge base module stores structured and unstructured domain knowledge and supports information retrieval during the dialogue process. The system organizes knowledge using vector-based retrieval and achieves efficient querying through semantic matching [15]. Figure 5 shows the knowledge base construction process.

Table II shows the management functions of the knowledge base.

TABLE II
KNOWLEDGE BASE MANAGEMENT FUNCTIONS

Function	Description
Knowledge Import	Import structured and unstructured knowledge into the knowledge base
Knowledge Update	Regularly update and maintain knowledge to ensure its accuracy and timeliness
Knowledge Retrieval	Retrieve relevant knowledge from the knowledge base based on user questions to support dialogue generation

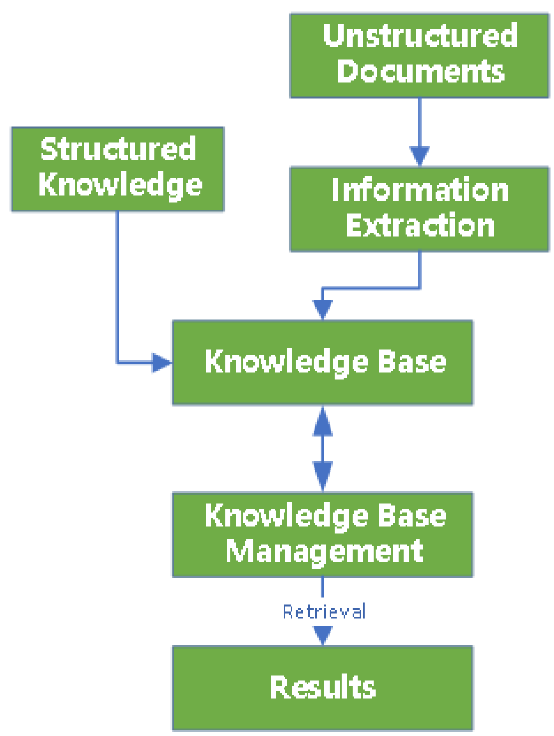

Fig. 5. Knowledge Base Construction Process

V. System Implementation and Evaluation

A. Experimental Environment and Datasets

The experiments were conducted on a GPU server with the hardware configuration shown in Table III. The system was implemented using PyTorch and loaded pre-trained models using the HuggingFace Transformers library.

TABLE III
Experimental Environment Parameters

Hardware/Software	Configuration
CPU	Intel Xeon Gold 6278C @ 2.60GHz
GPU	NVIDIA A100 80GB * 8
Memory	512 GB
Operating System	CentOS 7
Deep Learning Framework	PyTorch 1.11
NLP Library	HuggingFace Transformers 4.20

The following datasets were used in the experiments:

Pre-training dataset: Common Crawl Chinese corpus, a total of 100G tokens.

Fine-tuning dataset: Dialogue data in a financial customer service scenario, a total of 1 million dialogue samples.

Evaluation dataset: 10,000 real user inquiries and reference responses were manually annotated as the evaluation set.

B. Large Model Training and Evaluation

The large model adopts the GPT-2 model structure and performs self-supervised pre-training on the pre-training corpus. After pre-training, the model is fine-tuned on customer service data. Table IV shows the model hyperparameter settings.

TABLE IV
Model Hyperparameter Settings

Hyperparameter	Value
Model Structure	GPT-2 (L=24, H=1024, A=16)
Batch Size	64
Learning Rate	1e-4
Optimizer	Adam
Pre-training Steps	1,000,000
Fine-tuning Steps	100,000

We compared the performance of models of different sizes on the evaluation set using BLEU and ROUGE as evaluation metrics. Table V shows the evaluation results of different models. It can be seen that as the model size increases, the generation quality of the model continuously improves. The GPT2-Large model achieves the best performance, with its generated responses highly similar to the reference answers.

TABLE V
Performance Comparison of Different Models

Model	BLEU-4	ROUGE-L
GPT2-Small	20.31	35.67
GPT2-Medium	23.54	38.29
GPT2-Large	26.72	41.83

Figure 6 shows the loss curve of the GPT2-Large model fine-tuned on customer service data. It can be seen that the model converges quickly, and the loss tends to stabilize after 20,000 training steps.

Fig. 6. GPT2-Large Fine-tuning Loss Curve on Customer Service Data

C. Overall Evaluation of Intelligent Customer Service System

We conducted a comprehensive performance evaluation and user experience evaluation of the intelligent customer service system. In terms of performance evaluation, we collected statistics on indicators such as the system's response latency and daily user service volume. Table VI shows the system's performance statistics.

TABLE VI
System Performance Metrics

Metric	Value
Average First Response Latency	335 ms
Average Conversation Duration	3.12 min
Daily User Service Volume	12,350
Peak QPS	23

In terms of user experience evaluation, we invited 50 real users to interact with the system and fill out a satisfaction survey questionnaire. The survey scored the system (on a 5-point scale) from dimensions such as dialogue fluency, response relevance, and issue resolution. Table VII shows the manual evaluation results.

TABLE VII
Manual Evaluation Results

Evaluation Dimension	Average Score
Dialogue Fluency	4.32
Response Relevance	4.45
Issue Resolution	4.27
Overall Satisfaction	4.41

The evaluation results show that the intelligent customer service system based on large models can provide a fluent and natural dialogue experience, accurately understand user intent and provide relevant responses, effectively solving most user inquiries with high overall satisfaction.

Below is a typical dialogue example of the intelligent customer service system:

User: I recently took out a loan and would like to inquire about repayment matters.

Customer Service: Okay, may I ask what type of loan you took out? We support multiple repayment methods such as equal principal and interest, equal principal, and interest before principal.

User: I took out a personal credit loan for 200,000 yuan with a term of 3 years.

Customer Service: Based on the information you provided, you can choose the equal principal and interest or equal principal repayment method. Equal principal and interest means repaying the same amount of loan (including principal and interest) each month, with stable monthly payments but higher total interest paid; equal principal means repaying the same principal and interest on the remaining loan each month, with monthly payments decreasing each month and lower total interest paid. It is recommended that you choose a suitable repayment method based on your own situation.

User: I'll go with equal principal and interest, so the monthly repayment pressure will be smaller. About how much do I need to repay each month then?

Customer Service: Okay, you have chosen the equal principal and interest repayment method. Taking the 3-year LPR of 4.65

User: I see, repaying this amount each month is still acceptable.

Customer Service: Thank you for your inquiry. If you have any other loan-related questions, feel free to contact us at any time. I wish you a happy life. Goodbye!

The above case demonstrates the responses generated by the intelligent customer service system based on large models. It can be seen that the system can provide detailed and professional answers to users' specific questions, and the communication is efficient and friendly.

VI. Summary and Outlook

A. Research Summary

This paper explores the design of an intelligent customer service system based on large models. With large models at its core, the system achieves high-quality customer service dialogue generation through modular design and self-supervised learning. Experiments show that the system can significantly improve the level of customer service automation and service quality, providing enterprises with a brand-new customer service solution.

The main contributions of this paper include:

1) Proposing an intelligent customer service system architecture based on large models, adopting a modular design that is easy to extend and maintain.
2) Exploring large model fine-tuning methods combined with domain knowledge to achieve dialogue generation in specific domains.
3) Conducting comprehensive evaluations of the system on real-world scenario datasets, demonstrating its effectiveness and practicality.

B. Future Research Outlook

1) Introduce incremental learning to enable the system to learn new knowledge online and continuously improve dialogue capabilities.
2) Explore personalized dialogue generation techniques to provide customized services for different users.
3) Integrate emotion computing technology to enable the system to perceive user emotions and provide more considerate services.
4) Conduct human-machine collaboration research to build a new model of customer service with human-machine complementarity.

Intelligent customer service is an important scenario for artificial intelligence technology to empower enterprises. With the continuous development of technologies such as large models, intelligent customer service systems will become more and more powerful, creating greater value for enterprises and users. Let us work together to promote the innovative development of intelligent customer service.

References

[1] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer Learning*. Cambridge University Press, 2022.

[2] J.Wei et al., "Chain of thought prompting elicits reasoning in large language models," 2023.

[3] R.Thoppilan et al., "Lamda: Language models for dialog applications," *CoRR*, vol. abs/2201.08239, 2022.

[4] M. Nuruzzaman and O. K. Hussain, "A survey on chatbot implementation in customer service industry through deep neural networks," in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, 2018, pp. 54–61.

[5] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "Superagent: A customer service chatbot for e-commerce websites," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 97–102.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[8] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[10] L.Ouyang et al., "Training language models to follow instructions with human feedback," 2022.

[11] Y. Wang, C. Li, H. Huang, X. Sun, and Q. Wu, "A survey on dialogue summarization: Recent advances and new frontiers," 2021.

[12] Y.Zhang et al., "Dialogpt: Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 270–278.

[13] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[14] D.Adiwardana et al., "Towards a human-like open-domain chatbot," *CoRR*, vol. abs/2001.09977, 2020.

[15] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," *Foundations and Trends in Information Retrieval*, vol. 13, no. 2-3, pp. 127–298, 2019.

[16] Y.Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022.